

A Linear Model for Interval-valued Data

Yan Sun^{*1} and Dan Ralescu^{†2}

¹Department of Mathematics & Statistics, Utah State University, 3900 Old Main Hill,
Logan, Utah 84322

²Department of Mathematical Sciences, University of Cincinnati, 2600 Clifton Avenue,
Cincinnati, Ohio 45220

Abstract

Interval-valued linear regression has been investigated for some time. One of the critical issues is optimizing the balance between model flexibility and interpretability. This paper proposes a linear model for interval-valued data based on the affine operators in the cone $C = \{(x, y) \in \mathbb{R}^2 | x \leq y\}$. The resulting new model is shown to have improved flexibility over typical models in the literature, while maintaining a good interpretability. The least squares (LS) estimators of the model parameters are provided in a simple explicit form, which possesses a series of nice properties. Further investigations into the LS estimators shed light on the positive restrictions of a subset of the parameters and their implications on the model validity. A simulation study is presented that supports the theoretical findings. An application to a real data set is also provided to demonstrate the applicability of our model.

1 Introduction

Recently there has been an increasing interest in the linear regression for interval-valued data. See Diamond(1990), Körner and Näther (1998) , Gil et al. (2002, 2007), Manski and Tamer (2002), Carvalho et al. (2004), Billard (2007), González-Rodríguez et al. (2007), Lima Neto and De Carvalho (2008, 2010), Blanco-Fernández et al. (2011), Cattaneo and Wiencierz (2012), for a partial list of references. Existing models have been developed mainly in two directions. In the first direction, separate point-valued linear regression models are fitted to the center and range (or the lower and upper bounds), respectively, treating the intervals essentially as bivariate vectors. Examples belonging to this category include the center method by Billard and Diday (2000), the MinMax method by Billard and Diday (2002), the (constrained) center and range method by Lima Neto and De Carvalho (2008, 2010), and the model M by Blanco-Fernández et al. (2011). The second direction is to view the intervals as subsets in \mathbb{R} and study their linear relationship in the framework of random sets. Investigations along this direction include Diamond (1990), Gil et al. (2001, 2002), Gil et al. (2007), González-Rodríguez (2007), and Sun and Li (2014), among others. In this paper, we propose a new linear model for interval-valued data that aims at connecting the two directions and achieving improved flexibility.

To facilitate our presentation, let us give a brief introduction on the theoretical framework of random sets. Let (Ω, \mathcal{L}, P) be a probability space. Denote by $\mathcal{K}(\mathbb{R}^d)$ or \mathcal{K} the collection of all non-empty compact subsets of \mathbb{R}^d . In the space \mathcal{K} , a linear structure is defined by Minkowski addition and scalar multiplication, i.e.,

$$A + B = \{a + b : a \in A, b \in B\}, \quad \lambda A = \{\lambda a : a \in A\}, \quad (1)$$

$\forall A, B \in \mathcal{K}$ and $\lambda \in \mathbb{R}$. A random compact set is a Borel measurable function $A : \Omega \rightarrow \mathcal{K}$, \mathcal{K} being equipped with the Borel σ -algebra induced by the Hausdorff metric. For each $X \in \mathcal{K}(\mathbb{R}^d)$, the function defined on the unit sphere S^{d-1} :

$$s_X(u) = \sup_{x \in X} \langle u, x \rangle, \quad \forall u \in S^{d-1}$$

^{*}yan.sun@usu.edu

[†]dan.ralescu@uc.edu

is called the support function of X . If $A(\omega)$ is convex almost surely, then A is called a random compact convex set. (See Molchanov 2005, p.21, p.102.) The collection of all compact convex subsets of \mathbb{R}^d is denoted by $\mathcal{K}_C(\mathbb{R}^d)$ or \mathcal{K}_C . Especially, when $d = 1$, $\mathcal{K}_C(\mathbb{R})$ contains all the non-empty bounded closed intervals in \mathbb{R} . A measurable function $X : \Omega \rightarrow \mathcal{K}_C(\mathbb{R})$ is called a random interval. Much of the random sets theory has focused on compact convex sets (see, e.g., Artstein and Vitale (1975), Aumann (1965), and Lyashenko (1982, 1983)). Let \mathcal{S} be the space of support functions of all non-empty compact convex subsets in \mathcal{K}_C . Then, \mathcal{S} is a Banach space equipped with the L_2 metric

$$\|s_X(u)\|_2 = \left[d \int_{S^{d-1}} |s_X(u)|^2 \mu(du) \right]^{\frac{1}{2}},$$

where μ is the normalized Lebesgue measure on S^{d-1} . According to various embedding theorems (see Rådström 1952; Hörmander 1954), \mathcal{K}_C can be embedded isometrically into the Banach space $C(S)$ of continuous functions on S^{d-1} , and \mathcal{S} is the image of \mathcal{K}_C into $C(S)$. Therefore, $\delta(X, Y) := \|s_X - s_Y\|_2, \forall X, Y \in \mathcal{K}_C$, defines an L_2 metric on \mathcal{K}_C .

The central idea to constructing linear models in the random sets framework is to minimize the distance $\delta(Y, E(Y|X))$ on the data, where X, Y are random intervals and $E(Y|X)$ is a linear function of X in the sense of (1). Such models have very nice mathematical interpretations, but the restriction to the space $\mathcal{K}_C(\mathbb{R})$ unfortunately results in a reduced flexibility from practical point of view. Notice that

$$\begin{aligned} (aX + b)^C &= aX^C + b, \\ (aX + b)^R &= |a|X^R. \end{aligned}$$

This implies that the slope parameters of the corresponding linear model for the center (C) and range (R) must be the same in absolute value (see, e.g., Gil et al. (2002) and Sun and Li (2014)). Such a restriction is usually relaxed in the models developed for the center and range separately. Those models typically treat an interval as a vector in the Euclidean space \mathbb{R}^2 and minimize the Euclidean distance $\|Y - E(Y|X)\|$ on the data. However, this approach is slightly problematic in that once the intervals are represented by vectors in \mathbb{R}^2 , they should be modeled as such, as opposed to being broken down to the centers and ranges separately. Particularly, a linear model in \mathbb{R}^2 in general takes on the form

$$Y = AX + \mathbf{b} + \boldsymbol{\epsilon}, \tag{2}$$

where A is a 2×2 coefficient matrix, \mathbf{b} is a 2×1 intercept vector, and $\boldsymbol{\epsilon}$ is a 2×1 error vector. There is no reason to separate the two coordinates of X by forcing A to be diagonal. This problem makes the bivariate types of models hard to interpret both in $\mathcal{K}_C(\mathbb{R})$ and in \mathbb{R}^2 .

Our main contribution in this paper is to generalize the bivariate types of models from the literature (i.e., models in the first research direction by the preceding discussion) to the form (2), which is accomplished by embedding the space $\mathcal{K}_C(\mathbb{R})$ into \mathbb{R}^2 , and more precisely, into the cone $\mathcal{C} = \{(x, y) \in \mathbb{R}^2 | x \leq y\}$. As such, our proposed new linear model has generally improved flexibility over the existing models in both directions. It is also well interpretable in \mathcal{C} due to the embedding. We extend the univariate model to the multiple case and derive the matrix form of the general multivariate model. The least squares (LS) estimates for the model parameters are provided in matrix form, from which a series of properties are derived. Furthermore, we give explicit analytical LS solutions for the positive parameters, which shed light on the behaviors of the LS estimators in connection with the positive restriction and the model validity. Simulation studies are carried out that produce consistent results with our theoretical findings. Finally, an application to a real data set is presented to demonstrate the applicability of our model.

The rest of the paper is organized as follows. Section 2 formally introduces our model and discusses the associated model properties. The LS estimators and their properties are presented in Section 3, followed by a rigorous discussion on the estimation of the positive parameters in Section 4. Simulation studies are reported in Section 5, and the real data application is presented in Section 6. We give concluding remarks in Section 7. Technical proofs are collected in the Appendix.

2 The linear model

2.1 The affine operator in \mathcal{C} and the univariate model in $\mathcal{K}_{\mathcal{C}}(\mathbb{R})$

Assume observing an i.i.d. random sample of paired intervals $X_i = [X_i^L, X_i^U]$, $Y_i = [Y_i^L, Y_i^U]$, $i = 1, \dots, n$, where X_i^L , Y_i^L and X_i^U , Y_i^U are the lower and upper bounds of X_i and Y_i , respectively. Alternatively, the interval X_i can also be represented by its center X_i^C and range X_i^R as

$$\begin{aligned} X_i^C &= (X_i^U + X_i^L) / 2, \\ X_i^R &= X_i^U - X_i^L, \end{aligned}$$

and similarly for Y_i . The δ -metric in the space $\mathcal{K}_{\mathcal{C}}(\mathbb{R})$ is given by

$$\delta(X_1, X_2) = \sqrt{\frac{1}{2} (X_1^L - X_2^L)^2 + \frac{1}{2} (X_1^U - X_2^U)^2}.$$

This suggests that the metric space $(\mathcal{K}_{\mathcal{C}}(\mathbb{R}), \delta)$ can be embedded isometrically into the cone $\mathcal{C} = \{(x, y) \in \mathbb{R}^2 | x \leq y\}$ equipped with the Euclidean metric. Therefore, we consider each interval $X = [X^L, X^U] \in \mathcal{K}_{\mathcal{C}}(\mathbb{R})$ to be represented by the point $(X^L, X^U) \in \mathcal{C}$.

From the preceding discussion of embedding, we propose to construct a linear model in $\mathcal{K}_{\mathcal{C}}(\mathbb{R})$ based on the affine operator in \mathcal{C} , i.e., affine operator $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ satisfying $T(\mathcal{C}) \subseteq \mathcal{C}$. Obviously, such affine operators are represented by

$$T\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \alpha & \beta \\ \alpha - \gamma & \beta + \gamma \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \eta \\ \eta + \theta \end{bmatrix},$$

with $\alpha, \beta, \eta \in \mathbb{R}$ and $\gamma, \theta \geq 0$. This leads us to propose the following univariate linear model

$$Y_i^L = \alpha X_i^L + \beta X_i^U + \eta + \epsilon_i^L, \quad (3)$$

$$Y_i^U = (\alpha - \gamma) X_i^L + (\beta + \gamma) X_i^U + \eta + \theta + \epsilon_i^U, \quad (4)$$

where $\alpha, \beta, \eta \in \mathbb{R}, \gamma, \theta \geq 0$ are coefficients, and $\{\epsilon_i^L, \epsilon_i^U\}$ are i.i.d. zero mean random variables with variance $\sigma^2 > 0$, $i = 1, \dots, n$.

2.2 Collinearity preservation

The most important property of affine transformation is that it preserves collinearity. This in the cone \mathcal{C} means that points lying on a ray are still on a ray after transformation. Precisely, the operator T maps the ray

$$y = ax + b, \quad y \geq x$$

into another ray

$$T(y) = \left[1 + \frac{\gamma(a-1)}{\alpha + \beta a}\right] T(x) + \gamma b + \theta - \frac{\gamma(a-1)}{\alpha + \beta a} (\beta b + \eta), \quad T(y) \geq T(x).$$

Figure 1 gives an illustration of this effect. Considering the equivalence of the point $(X^L, X^U) \in \mathcal{C}$ and the interval $[X^L, X^U] \in \mathcal{K}_{\mathcal{C}}(\mathbb{R})$, we define a collection of intervals $\{[X_i^L, X_i^U] : i \in I\}$ to be collinear if their representations $\{(X_i^L, X_i^U) : i \in I\}$ in \mathcal{C} are on a ray.

Definition 1. A collection of intervals $\{[X_i^L, X_i^U] : i \in I\}$ are said to be collinear if they satisfy the equation

$$X_i^U = aX_i^L + b, \quad (5)$$

where $X_i^L \geq -\frac{b}{a-1}$ if $a > 1$, $X_i^U \leq -\frac{b}{a-1}$ if $a < 1$, and $b \geq 0$ if $a = 1$.

It is easily seen that equation (5) can be equivalently expressed as

$$X^R = \frac{2}{a+1} [(a-1)X^C + b].$$

So, we can also define collinearity in terms of the center and range of the interval.

Definition 2. *The collinearity of a collection of intervals $\{[X_i^L, X_i^U] : i \in I\}$ is equivalently defined by*

$$X_i^R = cX_i^C + d, \quad (6)$$

where $X_i^C \geq -\frac{d}{c}$ if $c > 0$, $X_i^C \leq \frac{d}{c}$ if $c < 0$, and $d \geq 0$ if $c = 0$.

From these two definitions, collinearity of intervals essentially means that the upper bound changes linearly with the lower bound, or equivalently, the range changes linearly with the center. For example, it is a common situation in practice that a larger center is associated with a wider range. When this relationship is linear, the corresponding intervals are considered collinear, and such a characteristic gets preserved under the operator T . Figure 2 provides a visualization of this property. In terms of modeling, if an interval-valued data $x_i = [x_i^L, x_i^U]$, $y_i = [y_i^L, y_i^U]$, $i = 1, \dots, n$, follows our model (3)-(4), then for x_i 's that are collinear, their associated y_i 's are also collinear.

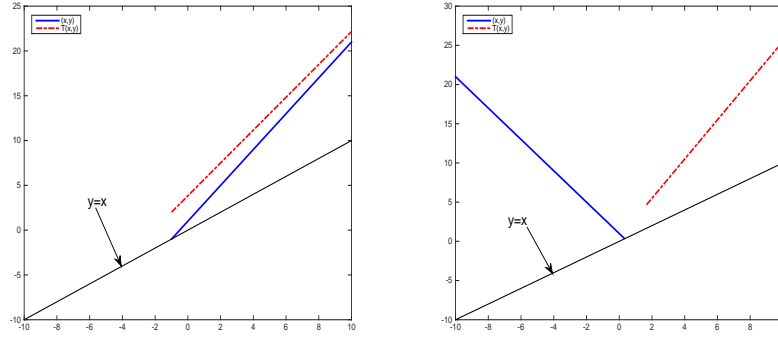


Figure 1: A graphical illustration of the affine transformation in \mathcal{C} , which is above the line $y = x$ in \mathbb{R}^2 . The solid line is a ray $y = ax + b, y \geq x$ in \mathcal{C} , and the dash-dotted line is its image by T with the parameters $\alpha = -2, \beta = 4, \gamma = 5, \eta = 1, \theta = 3$. Left: $a = 2, b = 1$; Right: $a = -2, b = 1$.

2.3 Comparison to other models

As we mentioned in the introduction, our model has systematically improved flexibility over typical models in the literature. In this section, we compare our univariate model (3)-(4) to two popular models to gain more insight into this. Consider the M model proposed by Blanco-Fernández et al. (2011), and the constrained center and range method (CCRM) by Lima Neto and De Carvalho (2010). The M model is specified as

$$Y_i^C = \alpha X_i^C + \gamma + \text{mid}\epsilon_i, \quad (7)$$

$$Y_i^R = |\beta|X_i^R + \text{spr}\epsilon_i, \quad (8)$$

where $\text{mid}\epsilon_i$ and $\text{spr}\epsilon_i$ are center and range of the interval-valued random error ϵ_i , respectively. $\text{mid}\epsilon_i$ is assume to be a centered random variable and $\text{spr}\epsilon_i$ is assumed to be a positive random variable. On the other hand, the model of CCRM is defined as

$$Y_i^C = \beta_0^C + \beta_1^C X_i^C + \epsilon_i^C, \quad (9)$$

$$Y_i^R = \beta_0^R + \beta_1^R X_i^R + \epsilon_i^R, \quad (10)$$

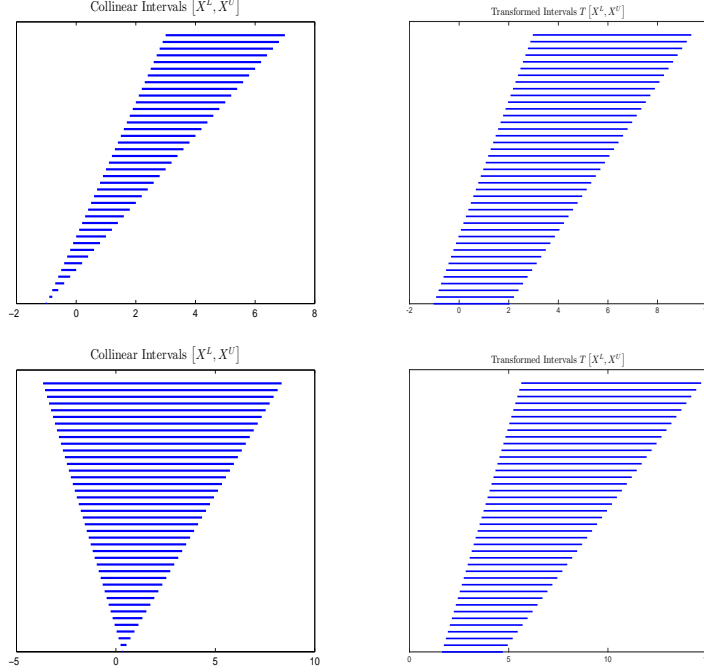


Figure 2: Visualization of collinear intervals. Each interval is displayed as a horizontal line segments, and the intervals are elevated proportionally in order to be displayed in one plot. The left two plots are collinear intervals that satisfy equation (5) with (top) $a = 2, b = 1$ and (bottom) $a = -2, b = 1$. The right two plots are their corresponding images by T with parameters $\alpha = -2, \beta = 4, \gamma = 5, \eta = 1, \theta = 3$.

where ϵ_i^C and ϵ_i^R are both centered random variables without any geometric interpretations. The two coefficients β_0^R, β_1^R in the range regression equation are both restricted to be positive to ensure the positiveness of Y_i^R . It is easy to see that these two models are essentially equivalent with $\epsilon_i^C = \text{mid}\epsilon_i$, $\beta_0^R = E(\text{spre}\epsilon_i)$, and $\epsilon_i^R = \text{spre}\epsilon_i - E(\text{spre}\epsilon_i)$. Rewriting equations (9)-(10) in terms of the lower and upper bounds, the model of CCRM is equivalently represented as

$$Y_i^L = \beta_0^C - \frac{1}{2}\beta_0^R + \frac{1}{2}(\beta_1^C + \beta_1^R)X_i^L + \frac{1}{2}(\beta_1^C - \beta_1^R)X_i^U + \epsilon_i^C - \frac{1}{2}\epsilon_i^R, \quad (11)$$

$$Y_i^U = \beta_0^C + \frac{1}{2}\beta_0^R + \frac{1}{2}(\beta_1^C - \beta_1^R)X_i^L + \frac{1}{2}(\beta_1^C + \beta_1^R)X_i^U + \epsilon_i^C + \frac{1}{2}\epsilon_i^R, \quad (12)$$

where $\beta_0^R > 0, \beta_1^R > 0$. This compared to our model (3)-(4) is a reduced form with the restrictions $\alpha = \beta + \gamma$. So our model has one extra degree of freedom, which will drastically expand the model flexibility. The CCRM is extended to the multiple case, from which the advantage of our general model introduced in the following gets multiplied. We will elaborate more on this in the simulation and real data application sections.

2.4 Matrix form of the general model

Consider the general case involving the outcome interval $Y_i = [Y_i^L, Y_i^U]$ and p interval-valued predictors $X_{j,i} = [X_{j,i}^L, X_{j,i}^U]$, $i = 1, \dots, n$; $j = 1, \dots, p$. To model Y_i by a linear transformation of $\{X_{j,i} : j = 1, \dots, p\}$, we extend the univariate model (3)-(4) to the following form:

$$Y_i^L = \sum_{j=1}^p (\alpha_j X_{j,i}^L + \beta_j X_{j,i}^U) + \eta + \epsilon_i^L, \quad (13)$$

$$Y_i^U = \sum_{j=1}^p [(\alpha_j - \gamma_j) X_{j,i}^L + (\beta_j + \gamma_j) X_{j,i}^U] + \eta + \theta + \epsilon_i^U, \quad (14)$$

with $\gamma_j \geq 0$, $E(\epsilon_i^U) = E(\epsilon_i^L) = 0$, and $\text{Var}(\epsilon_i^U) = \text{Var}(\epsilon_i^L) = \sigma^2 > 0$, $\theta \geq 0$, for $i = 1, \dots, n$ and $j = 1, \dots, p$.

Define

$$\mathbf{Y}^L = \begin{bmatrix} Y_1^L \\ Y_2^L \\ \vdots \\ \vdots \\ Y_n^L \end{bmatrix}, \quad \mathbf{X}_1 = \begin{bmatrix} 1 & X_{1,1}^L & X_{1,1}^U & X_{2,1}^L & X_{2,1}^U & \cdots & X_{p,1}^L & X_{p,1}^U \\ 1 & X_{1,2}^L & X_{1,2}^U & X_{2,2}^L & X_{2,2}^U & \cdots & X_{p,2}^L & X_{p,2}^U \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & X_{1,n}^L & X_{1,n}^U & X_{2,n}^L & X_{2,n}^U & \cdots & X_{p,n}^L & X_{p,n}^U \end{bmatrix},$$

and

$$\boldsymbol{\beta}_1 = \begin{bmatrix} \eta \\ \alpha_1 \\ \beta_1 \\ \vdots \\ \vdots \\ \alpha_p \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon}^L = \begin{bmatrix} \epsilon_1^L \\ \epsilon_2^L \\ \vdots \\ \vdots \\ \vdots \\ \epsilon_n^L \end{bmatrix}.$$

Then equation (13) is expressed as

$$\mathbf{Y}^L = \mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^L.$$

Define

$$\mathbf{Y}^U = \begin{bmatrix} Y_1^U \\ Y_2^U \\ \vdots \\ \vdots \\ Y_n^U \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} 1 & X_{1,1}^R & X_{2,1}^R & \cdots & X_{p,1}^R \\ 1 & X_{1,2}^R & X_{2,2}^R & \cdots & X_{p,2}^R \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1,n}^R & X_{2,n}^R & \cdots & X_{p,n}^R \end{bmatrix},$$

and

$$\boldsymbol{\beta}_2 = \begin{bmatrix} \theta \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \vdots \\ \gamma_p \end{bmatrix}, \quad \boldsymbol{\epsilon}^U = \begin{bmatrix} \epsilon_1^U \\ \epsilon_2^U \\ \vdots \\ \vdots \\ \vdots \\ \epsilon_n^U \end{bmatrix}.$$

Then equation (14) is rewritten as

$$\mathbf{Y}^U = [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{\epsilon}^U.$$

To jointly express the model, define

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}^L \\ \mathbf{Y}^U \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}^L \\ \boldsymbol{\epsilon}^U \end{bmatrix}.$$

Then, the general model (13)-(14), can be written in the matrix form

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}. \quad (15)$$

3 Least squares estimation

We define our least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ as the minimizer of the sum of squared lower and upper bound errors. Namely,

$$\hat{\boldsymbol{\beta}} = \arg \min \left\{ \sum_{i=1}^n \left[\left(Y_i^L - \hat{Y}_i^L \right)^2 + \left(Y_i^U - \hat{Y}_i^U \right)^2 \right] \right\}, \quad (16)$$

where

$$\hat{Y}_i^L = \sum_{j=1}^p (\alpha_j X_{j,i}^L + \beta_j X_{j,i}^U) + \eta, \quad (17)$$

$$\hat{Y}_i^U = \sum_{j=1}^p [(\alpha_j - \gamma_j) X_{j,i}^L + (\beta_j + \gamma_j) X_{j,i}^U] + \eta + \theta. \quad (18)$$

This is equivalent to minimizing the sum of squared δ -distance in the metric space $(\mathcal{K}_C(\mathbb{R}), \delta)$. Theorem 1 gives the explicit analytical expression of $\hat{\beta}$ in matrix form.

Theorem 1. *Consider the linear model (13)-(14), or equivalently, its matrix form (15). If the data matrix X has full rank, then the least squares estimate $\hat{\beta}$ defined in (16) is given by*

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (19)$$

Departing from its matrix form, a series of nice properties of $\hat{\beta}$ follows immediately from the classical theory of linear models. (See, e.g., Seber (1997).) We summarize them in the following corollaries.

Corollary 1. *$\hat{\beta}$ in Theorem 1 is unbiased.*

Corollary 2. *$\hat{\beta}$ in Theorem 1 is consistent.*

Corollary 3. *The variance-covariance matrix of $\hat{\beta}$ is*

$$\text{Cov}(\hat{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2. \quad (20)$$

Corollary 4. *An unbiased estimator of σ^2 is given by*

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})^T (\mathbf{Y} - \mathbf{X}\hat{\beta})}{2n - 3p - 2}. \quad (21)$$

4 Positive restrictions

The model setting requires that $\gamma_j \geq 0$, $\theta \geq 0$, $j = 1, \dots, p$. However, $\hat{\beta}$ given in (19) does not automatically guarantee these conditions. In this section, we thoroughly discuss these positive restrictions for the least squares estimation and their implications on the model fitting. We begin by making a few notations and assumptions.

Notation 1. *Denote by X_k^V and Y^V the random variables from which $\{X_{k,i}^V\}_{i=1}^n$ and $\{Y_i^V\}_{i=1}^n$ are samples, respectively, where $k = 1, \dots, p$ and $V \in \{L, U\}$.*

Notation 2. *Denote by*

$$S_{k,j} = \frac{1}{n} \sum_{i=1}^n X_{k,i}^R X_{j,i}^R - \left(\frac{1}{n} \sum_{i=1}^n X_{k,i}^R \right) \left(\frac{1}{n} \sum_{i=1}^n X_{j,i}^R \right)$$

the sample covariance of X_k^R and X_j^R , $k, j = 1, \dots, p$. Similarly, denote by

$$S_k = \frac{1}{n} \sum_{i=1}^n X_{k,i}^R Y_{k,i}^R - \left(\frac{1}{n} \sum_{i=1}^n X_{k,i}^R \right) \left(\frac{1}{n} \sum_{i=1}^n Y_i^R \right)$$

the sample covariance of X_k^R and Y^R , $k = 1, \dots, p$.

Assumption 1. The ranges of the predictors $\{X_j^R, j = 1, \dots, p\}$ are mutually uncorrelated.

Assumption 2. The range of each predictor X_j^R is empirically positively correlated with the range of the outcome Y^R , i.e., $S_j > 0$ for $j = 1, \dots, p$.

From the model specification (13)-(14), it is seen that

$$Y_i^R = \sum_{j=1}^p \gamma_j X_{j,i}^R + \theta + \epsilon_i^R, \quad (22)$$

where $\epsilon_i^R = \epsilon_i^U - \epsilon_i^L$, $i = 1, \dots, n$. This immediately implies the following results, which give interpretations of the positive parameters θ and γ_j , $j = 1, \dots, p$.

Proposition 1. Assume model (13)-(14). Then,

$$1. \begin{bmatrix} \text{Cov}(X_1^R, Y^R) \\ \vdots \\ \text{Cov}(X_p^R, Y^R) \end{bmatrix} = \begin{bmatrix} \text{Cov}(X_1^R, X_1^R) & \cdot & \cdot & \cdot & \text{Cov}(X_1^R, X_p^R) \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \text{Cov}(X_p^R, X_1^R) & \cdot & \cdot & \cdot & \text{Cov}(X_p^R, X_p^R) \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_p \end{bmatrix};$$

$$2. \theta = E(Y^R) - \sum_{j=1}^p \gamma_j E(X_j^R).$$

It can be shown that the positive parameters $\{\theta, \gamma_j, j = 1, \dots, p\}$ are indeed estimated independently from the rest of the parameters $\{\alpha_j, \beta_j, j = 1, \dots, p\}$. We list their analytical LS solutions separately in the following theorem.

Theorem 2. Consider model (13)-(14). Define the sample variance-covariance matrix of $\{X_j^R, j = 1, \dots, p\}$ as

$$\Sigma_{\mathbf{X}^R} = \left[\frac{1}{n} \sum_{i=1}^n X_{k,i}^R X_{j,i}^R - \left(\frac{1}{n} \sum_{i=1}^n X_{k,i}^R \right) \left(\frac{1}{n} \sum_{i=1}^n X_{j,i}^R \right) \right]_{k,j=1}^p := [S_{k,j}]_{k,j=1}^p.$$

Additionally, denote by $\Sigma_{\mathbf{X}^R, \mathbf{Y}^R}$ the vector that contains the sample covariances of X_k^R and Y^R , $k = 1, \dots, p$, i.e.

$$\Sigma_{\mathbf{X}^R, \mathbf{Y}^R} = \left[\frac{1}{n} \sum_{i=1}^n X_{k,i}^R Y_{i,i}^R - \left(\frac{1}{n} \sum_{i=1}^n X_{k,i}^R \right) \left(\frac{1}{n} \sum_{i=1}^n Y_i^R \right) \right]_{k=1}^p := [S_k]_{k=1}^p.$$

Let $\Gamma = [\gamma_1, \dots, \gamma_p]^T$. Then, the LS estimator $\hat{\Gamma}$ is the solution of the linear system

$$\Sigma_{\mathbf{X}^R} \Gamma = \Sigma_{\mathbf{X}^R, \mathbf{Y}^R}, \quad (23)$$

and the LS estimator of θ is

$$\hat{\theta} = \overline{Y^R} - \sum_{j=1}^p \hat{\gamma}_j \overline{X_j^R}. \quad (24)$$

From (23)-(24) and in view of Proposition 1, we see that $\{\hat{\theta}, \hat{\gamma}_j, j = 1, \dots, p\}$ are essentially moment estimators of the underlying parameters, which are in fact strongly consistent. This also explains from another perspective the consistency shown in Corollary 2. In particular, an important interpretation of Theorem 2 is that if at least one of the positive parameters is estimated to be negative for a large sample size, it indicates that the underlying true parameter is negative with a high probability, and forcing the parameter to be positive may result in possible biases. We give a simplified example in the following Corollary to illustrate the implication of the positive restriction on $\{\gamma_j, j = 1, \dots, p\}$.

Corollary 5. Consider the univariate model (3)-(4). Let $\hat{\gamma}$ be the LS estimate and $\tilde{\gamma}$ be any constrained LS estimate of γ such that $\tilde{\gamma} \geq 0$. If $\hat{\gamma} < 0$, then

$$\sum_{i=1}^n (Y_i^R - \tilde{Y}_i^R)^2 \geq \sum_{i=1}^n (Y_i^R - \overline{Y^R})^2,$$

where \hat{Y}_i^R is the predicted value for Y_i^R based on the constrained LS estimates. The “=” holds if and only if $\tilde{\gamma} = 0$.

For the univariate model, if the LS estimate of γ is negative, forcing it to be positive will result in the model being worse than the constant model \bar{Y}^R for the range. Similar biases are expected for the multivariate cases too. Therefore, it is not recommended that a constrained optimization algorithm always be used to ensure positive estimates, if some of the LS estimates $\{\hat{\gamma}_j, j = 1, \dots, p\}$ are negative. At least a different model that accounts for the negative LS estimates should be considered as an alternative to the constrained linear model. In practice, it is often assumed that the predictors $\{X_j, j = 1, \dots, p\}$ are independent. We provide a sufficient condition under which the LS estimates $\{\hat{\gamma}_j, j = 1, \dots, p\}$ are positive with probability converging to one.

Corollary 6. *Under Assumption 1, $\hat{\gamma}_j > 0$ with probability going to one if Assumption 2 is met.*

Intuitively, under the circumstance of independent predictors, model (13)-(14) implies that

$$\text{Cov}(X_j^R, Y^R) > 0, \quad j = 1, \dots, p.$$

Consequently, for data that the model is appropriate for, the sample covariances $\{S_j, j = 1, \dots, p\}$ are positive almost surely, which by Theorem 2 is sufficient to ensure the positiveness of $\{\hat{\gamma}_j, j = 1, \dots, p\}$. Otherwise, the $\hat{\gamma}_j$'s can be negative, but that is essentially because one or more of the predictors are negatively correlated with the outcome in range and hence the model is not appropriate.

From the preceding discussion, if $\hat{\gamma}_j > 0, j = 1, \dots, p$, it means that the model fits the linear structure of the data very well. At this point, if $\hat{\theta} < 0$, it may not be worth forcing it to be positive using a constrained optimization, as that may bring unnecessary biases. The following theorem gives a guidance of judgment for such a situation.

Theorem 3. *Assume model (13)-(14), or its equivalent matrix form (15). Let*

$$\hat{Y}_i^R = \sum_{j=1}^p \gamma_j X_{j,i}^R + \theta \quad (25)$$

be the model predicted value for Y_i^R . Then,

$$P(\hat{Y}_i^R < 0) \leq \frac{\text{Var}(Y_i^R) - \text{Var}(\hat{Y}_i^R)}{(Y_i^R)^2} = \frac{2\sigma^2}{(Y_i^R)^2}. \quad (26)$$

Given a negative $\hat{\theta}$, it is possible to get negative predicts for Y^R . However, if the unexplained variance of Y^R is very small compared to the scale of $(Y^R)^2$, the chance to get a negative predict is tiny, and the rare cases of negative predict, if happened, can be rounded up to 0. In practice, the unexplained variance of Y^R is estimated by $2\hat{\sigma}^2$, which is then compared to the scale of $(Y^R)^2$ from the data to decide whether to stay with the negative unbiased LS estimate $\hat{\theta}$ or resort to a constrained LS estimate.

5 Simulation

We present a simulation study to demonstrate the empirical performance of the LS estimates and compare our model to some peer models in the literature. In particular, we consider the following four model configurations:

- I: $p=1$, $\eta, \alpha_1, \beta_1 \sim \text{Unif}(0, 4)$, $\theta, \gamma_1 \sim \text{Unif}(1, 3)$, and $\epsilon_i^L, \epsilon_i^U \sim \text{Unif}(0, \sigma^2)$ with $\sigma \sim \text{Unif}(2, 4)$, $i = 1, \dots, n$;
- II: $p=1$, $\eta, \alpha_1, \beta_1 \sim \text{Unif}(-4, 0)$, $\theta, \gamma_1 \sim \text{Unif}(1, 3)$, and $\epsilon_i^L, \epsilon_i^U \sim \text{Unif}(0, \sigma^2)$ with $\sigma \sim \text{Unif}(2, 4)$, $i = 1, \dots, n$;

- III: $p=3$, $\eta, \alpha_j, \beta_j \sim \text{Unif}(-4, 4)$, $\theta, \gamma_j \sim \text{Unif}(1, 3)$, $j = 1, 2, 3$, and $\epsilon_i^L, \epsilon_i^U \sim \text{Unif}(0, \sigma^2)$ with $\sigma \sim \text{Unif}(2, 4)$, $i = 1, \dots, n$.

The first two are univariate models, with positive and negative interval correlations between Y and X_1 , respectively. Figure 3 shows a plot of simulated data with $n = 100$ observations from each of the two models. The third one is a 3-dimensional model, with Y and X_j , $j = 1, 2, 3$, either positively or negatively correlated. A particular data with $n = 100$ observations simulated from this model is visualized in Figure 4, where it is seen that Y is positively correlated with both X_1 and X_2 , and negatively correlated with X_3 .

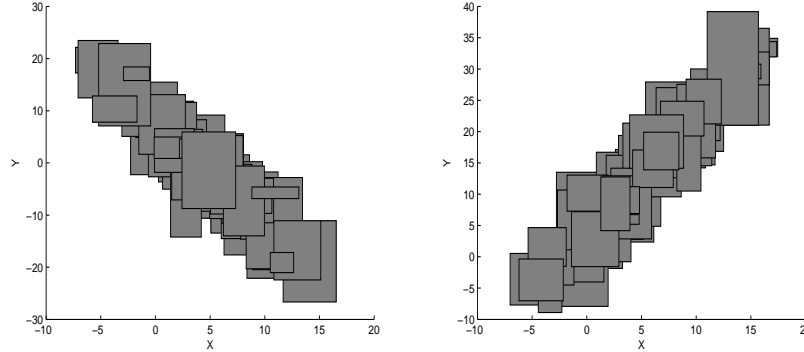


Figure 3: Plots of simulated data from model I and II, respectively, each with sample size $n = 100$.

To investigate the empirical performance of the LS estimation, we simulate 500 independent data from each of the three model configurations and calculate the LS estimates of the parameters for each simulated data. The results are summarized into Table 1. The mean relative error (MRE) for the estimated coefficient matrix $\hat{\beta}$ and variance of error $\hat{\sigma}^2$, given a fixed sample size n , are defined as

$$\text{MRE}(\hat{\beta}) = \frac{1}{500} \sum_{k=1}^{500} \frac{\|\hat{\beta}_k - \beta_k\|}{\|\beta_k\|},$$

where $\|\cdot\|$ denotes the Euclidean norm, and

$$\text{MRE}(\hat{\sigma}^2) = \frac{1}{500} \sum_{k=1}^{500} \frac{|\hat{\sigma}_k^2 - \sigma_k^2|}{\sigma_k^2},$$

respectively. We simulate observations for each X_i independently, so Assumption 1 is automatically satisfied. Assumption 2 is checked before we compute the LS estimates for the parameters for each simulated data. If it is satisfied, then $\hat{\beta}$ is calculated by (19), which according to Corollary 6 produces positive $\hat{\gamma}_j$, $j = 1, \dots, p$ with probability going to one. If otherwise Assumption 2 is violated, a constrained optimization algorithm is employed to calculate $\hat{\beta}$, with the constraints that $\hat{\gamma}_j \geq 0$, $j = 1, \dots, p$ and $\hat{\theta} \geq 0$. For this paper, we have used the Matlab function *fmincon.m* to compute the constrained LS estimates. Consistent to our theorems, we see that the MRE's for both $\hat{\beta}$ and $\hat{\sigma}^2$ converge to 0 as sample size increases. Especially, if the model really fits the data, which is the case for our simulation, the unconstrained LS estimate given in (19) is sufficient, without the need of a constrained optimization algorithm, with probability going to one.

Next, we carry out more delicate investigations into the parameter estimation with a particular model randomly generated from configuration III. The exact parameter values are listed in the second column of Table 2. We simulate a random sample of size $n = 300$ from this model and estimate the coefficient matrix β using the algorithm described in the preceding paragraph. The procedure is repeated for 500 times independently, and the mean estimates and mean variances are reported in columns 3 and 4, respectively. It is seen that the mean estimates are very close to the corresponding true values. The empirical variances of these estimates for the 500 repetitions are displayed in column 5, which are satisfactorily close to the calculated variances in column 4.

Table 1: Evaluation of the LS estimation for simulated data based on 500 independent repetitions.

	n	MRE ($\hat{\beta}$)	MRE ($\hat{\sigma}^2$)	Unconstrained	Constrained
Model I	100	0.4764	0.0982	496	4
	200	0.3806	0.0796	499	1
	300	0.3539	0.0726	500	0
	400	0.3386	0.0695	500	0
Model II	100	0.4468	0.0985	499	1
	200	0.3844	0.086	499	1
	300	0.3613	0.0764	500	0
	400	0.3331	0.0728	500	0
Model III	100	0.4581	0.0832	460	40
	200	0.3305	0.0546	473	27
	300	0.2666	0.0463	478	22
	400	0.2352	0.0391	481	19

Table 2: Parameter estimation for one particular model based on 500 independent repetitions.

Parameter	True Value	Mean Estimate	Estimated Variance	Empirical Variance
η	1.4932	1.4499	2.0048	1.8581
α_1	1.6419	1.6457	0.0520	0.0527
β_1	1.5542	1.5527	0.0521	0.0524
α_2	-1.8902	-1.9098	0.0521	0.0557
β_2	-3.2780	-3.2585	0.0521	0.0558
α_3	-2.4036	-2.3967	0.0518	0.0519
β_3	-1.8451	-1.8528	0.0518	0.0508
θ	1.7999	1.8149	3.8753	3.5164
γ_1	1.2086	1.2276	0.1033	0.1031
γ_2	2.5633	2.5347	0.1035	0.1112
γ_3	2.5436	2.5477	0.1028	0.0971

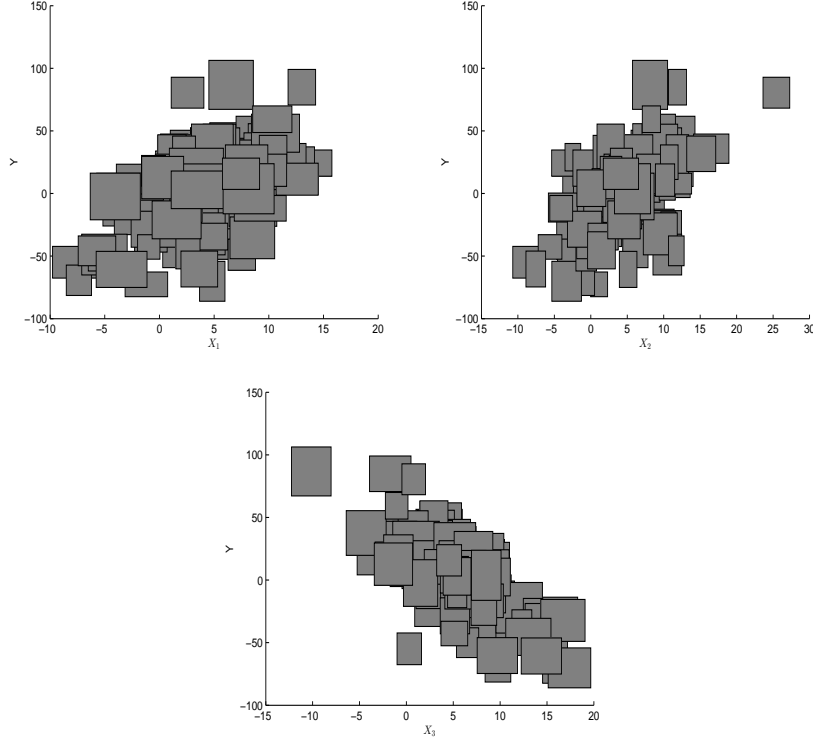


Figure 4: Plots of Y against X_1 , X_2 , and X_3 , respectively, of a simulated data from model III with sample size $n = 100$.

Finally, we compare our linear model to the M model by Blanco-Fernández et al. (2011) and the constrained center and range method (CCRM) by Lima Neto and De Carvalho (2010). We presented in Section 2.3 that these two models are essentially reduced forms of our model. Here we give empirical evidence based on their predicting performances. We simulate 500 independent samples from Model I, II, and III, with training sample size $n = 60, 100, 200, 300$, respectively. For each sample, we simulate another $n/4$ observations as the validation set. We use the mean squared error (MSE) of the center, radius (half-range), and the interval as a whole, for the validation set as our measures of predicting performance. Specifically, they are defined as

$$\begin{aligned} \text{MSEC} &= \frac{4}{n} \sum_{i=1}^{n/4} \left(\hat{Y}_i^c - Y_i^c \right)^2, \\ \text{MSER} &= \frac{4}{n} \sum_{i=1}^{n/4} \left(\hat{Y}_i^r - Y_i^r \right)^2, \\ \text{MSEI} &= \text{MSEC} + \text{MSER}. \end{aligned}$$

The LS solution for the parameters of the M model is calculated according to the formulas given in Blanco-Fernández et al. (2011). The CCRM is implemented using the R function *ccrm* in the *iRegression* package. The M model was only developed for the univariate case, so it is excluded in the multiple case (Model III). Numerical results for comparing the three methods are shown in Table 3. Just as we expected, the performance of our model is consistently significantly better than the other two models across different model configurations and sample sizes. Especially for Model III, the average MSEC of the CCRM is about 3 times bigger than that of our model, which results from the increased number of predictors. That is, the expanded flexibility of our model increases proportionally with the size of the model. The more predictors we include in the model, the more increased flexibility we have over the CCRM and the M model.

Table 3: Comparison of CCRM and our model for simulated data based on the average of 500 independent repetitions.

		M			CCRM			Our Linear Model		
	n	MSEC	MSER	MSEI	MSEC	MSER	MSEI	MSEC	MSER	MSEI
Model I	60	6.3398	4.6962	11.0361	6.2085	4.8585	11.067	4.8214	4.1395	8.961
	100	6.1048	4.4821	10.5869	6.1318	4.7239	10.8557	4.775	3.9936	8.7686
	200	6.1018	4.6074	10.7092	6.0296	4.5987	10.6283	4.8072	4.0166	8.8238
	300	5.991	4.494	10.4849	5.9705	4.7316	10.7021	4.6907	3.9587	8.6493
Model II	60	6.5056	4.4005	10.9062	6.4052	4.8404	11.2456	4.9052	4.0148	8.92
	100	6.178	4.567	10.745	6.0754	4.6265	10.7019	4.7378	3.9721	8.7098
	200	6.1372	4.6152	10.7525	6.0185	4.6246	10.6431	4.6802	3.9186	8.5988
	300	6.0026	4.4992	10.5018	5.9019	4.7009	10.6029	4.5837	3.9066	8.4903
Model III	60	-	-	-	14.1623	5.1865	19.3488	5.1949	4.9172	10.1122
	100	-	-	-	13.2387	4.8486	18.0873	5.0919	4.8285	9.9205
	200	-	-	-	13.3159	4.7531	18.069	4.7472	4.6946	9.4418
	300	-	-	-	13.1125	4.825	17.9375	4.6887	4.634	9.3227

6 A real data application

In this section, we apply our linear model to analyze an interval-valued climate data provided by the National Oceanic and Atmospheric Administration (NOAA) and publicly available. The data contains three variables. The outcome variable, which we denote by Y , is the average [minimum, maximum] temperature in July based on weather data collected from 1981 to 2010 by the NOAA National Climatic Data Center of the United States. The first predictor X_1 is the corresponding average temperature range in April. The second predictor X_2 is the [morning, afternoon] relative humidity in July averaged for the years 1961 to 1990. Relative humidity measures the actual amount of moisture in the air as a percentage of the maximum amount of moisture the air can hold, and it corresponds negatively to the temperature. All the three interval-valued variables are observed for 51 large US cities. By this analysis, we aim to model the summer (July) temperature by an affine function of the spring (April) temperature and the July relative humidity. We randomly split the full data into a training set of 40 observations and a validation set of 11 observations. Figure 5 plots Y against X_1 and X_2 , respectively, for the training set. It is checked that 1) the data matrix \mathbf{X} has full rank; 2) the sample correlation of X_1^R and X_2^R has a p-value greater than 0.05; 3) the sample correlations of X_i^R and Y^R , $i = 1, 2$, are both positive. So all the theoretical results we developed in Section 3 and 4 should apply. This means that with large probability we can get the LS estimates $\hat{\beta}$ simply by formula (19), without any constrained optimization algorithm. The LS estimates of the parameters are found to be

$$\begin{aligned}\eta &= 45.7740, \theta = 1.5419, \\ \alpha_1 &= 0.2069, \beta_1 = 0.3392, \gamma_1 = 0.8839, \\ \alpha_2 &= -0.0593, \beta_2 = -0.0948, \gamma_2 = 0.0234,\end{aligned}$$

and the estimated variance of residual according to Corollary 4 is

$$\hat{\sigma}^2 = 16.0454.$$

It follows that the fitted linear model is

$$\begin{aligned}Y^L &= 45.7740 + 0.2069X_1^L + 0.3392X_1^U - 0.0593X_2^L - 0.0948X_2^U + \epsilon^L, \\ Y^U &= 47.3159 - 0.6269X_1^L + 1.1731X_1^U - 0.0826X_2^L - 0.0714X_2^U + \epsilon^U,\end{aligned}$$

where ϵ^L, ϵ^U are i.i.d. random variables with mean 0 and variance 16.0454. For comparison purposes, we also fit a CCRM model to the data, which turns out to be

$$\begin{aligned} Y^C &= 57.2428 + 0.5768X_1^C - 0.1921X_2^C + \epsilon^C, \\ Y^R &= 1.5419 + 0.8339X_1^R + 0.0234X_2^R + \epsilon^R, \end{aligned}$$

where the random errors ϵ^C, ϵ^R have both means 0, and variances 19.7509 and 12.9017, respectively. The predicting performances on the validation set of both models are reported in table 4. Consistent with our theoretical analysis in Section 2.3 and our simulation study in Section 5, our linear model has much more flexibility than the existing reduced models such as CCRM, which leads to the much improved predicting performance even for a small data set as presented here.

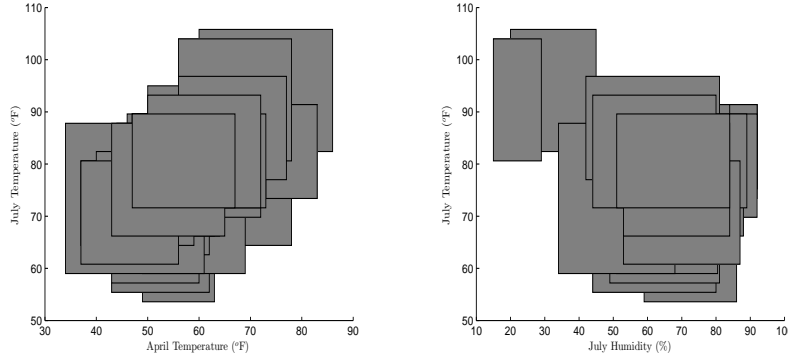


Figure 5: Left: plot of July temperature versus April temperature. Right: plot of July temperature versus July relative humidity.

Table 4: Predicting performance comparison of CCRM and our model for the real data.

	MSEC	MSER	MSEI
Our Linear Model	8.7484	4.1004	12.8488
CCRM	10.8631	4.1004	14.9634

7 Conclusion

We have introduced a linear model for interval-valued data based on the affine operators in the cone $\mathcal{C} = \{(x, y) \in \mathbb{R}^2 | x \leq y\}$. The new model is shown both theoretically and empirically to have improved flexibility over the existing models in the literature. We present the general model for multiple predictors in matrix form, from which the LS estimators of the model parameters are immediately derived with a series of nice properties from the classical theory of linear models. Some parameters have positive constraints, which we show are closely related to the intrinsic structure of the model. Therefore, it is not recommended to blindly force these parameters to be positive with a constrained optimization algorithm. Instead, it is better to let the data speak for itself by the unconstrained LS estimates and decide later whether to employ a constrained optimization algorithm or resort to a different model, according to the guideline we have provided in the paper.

A Proofs

A.1 Proof of Proposition 1

Proof. From (22), since ϵ^R and X_k^R are uncorrelated,

$$\text{Cov}(X_k^R, Y^R) = \text{Cov}\left(X_k^R, \sum_{j=1}^p \gamma_j X_j^R + \theta + \epsilon^R\right) = \sum_{j=1}^p \gamma_j \text{Cov}(X_k^R, X_j^R), \quad k = 1, \dots, p,$$

from which the first result follows. Taking expectations on both sides of (22) yields the second result. \square

A.2 Proof of Theorem 2

Proof. Differentiating $\sum_{i=1}^n \left[(Y_i^L - \hat{Y}_i^L)^2 + (Y_i^U - \hat{Y}_i^U)^2 \right]$ with respect to $\{\alpha_k, \beta_k, \gamma_k, \eta, \theta, k = 1, \dots, p\}$, we obtain the system of equations

$$\sum_{i=1}^n (Y_i^U - \hat{Y}_i^U) = 0, \quad (27)$$

$$\sum_{i=1}^n (Y_i^L - \hat{Y}_i^L) = 0, \quad (28)$$

$$\sum_{i=1}^n X_{k,i}^R (Y_i^U - \hat{Y}_i^U) = 0, \quad (29)$$

$$\sum_{i=1}^n X_{k,i}^R (Y_i^L - \hat{Y}_i^L) = 0, \quad (30)$$

$$\sum_{i=1}^n X_{k,i}^R \left[(Y_i^U - \hat{Y}_i^U) + (Y_i^L - \hat{Y}_i^L) \right] = 0, \quad (31)$$

$k = 1, \dots, p.$

Equations (27)-(28) yield

$$\sum_{i=1}^n Y_i^R = \sum_{j=1}^p \gamma_j \left(\sum_{i=1}^n X_{j,i}^R \right) + n\theta. \quad (32)$$

Meanwhile, equations (29)-(30) yield

$$\sum_{j=1}^p \gamma_j \left(\sum_{i=1}^n X_{k,i}^R X_{j,i}^R \right) + \theta \sum_{i=1}^n X_{k,i}^R = \sum_{i=1}^n X_{k,i}^R Y_i^R, \quad k = 1, \dots, p. \quad (33)$$

Plugging (32) into (33), we obtain

$$\begin{aligned} & \sum_{j=1}^p \gamma_j \left[\frac{1}{n} \sum_{i=1}^n X_{k,i}^R X_{j,i}^R - \left(\frac{1}{n} \sum_{i=1}^n X_{k,i}^R \right) \left(\frac{1}{n} \sum_{i=1}^n X_{j,i}^R \right) \right] \\ &= \frac{1}{n} \sum_{i=1}^n X_{k,i}^R Y_i^R - \left(\frac{1}{n} \sum_{i=1}^n X_{k,i}^R \right) \left(\frac{1}{n} \sum_{i=1}^n Y_i^R \right), \quad k = 1, \dots, p. \end{aligned} \quad (34)$$

Writing equations (34) in matrix form yields (23). (24) is obtained by plugging $\gamma_j, j = 1, \dots, p$ in equation (32). \square

A.3 Proof of Corollary 5

Proof. From Theorem 2, $\hat{\gamma} = S_1/S_{1,1} < 0$, where S_1 and $S_{1,1}$ are the sample covariance of X^R and Y^R , and the sample variance of X^R , respectively. Namely,

$$S_1 = \frac{1}{n} \sum_{i=1}^n (Y_i^R - \overline{Y^R}) (X_i^R - \overline{X^R}) < 0, \quad (35)$$

$$S_{1,1} = \frac{1}{n} \sum_{i=1}^n (X_i^R - \overline{X^R})^2 > 0. \quad (36)$$

Let $[\tilde{\gamma}, \tilde{\theta}]$ be the joint constrained LS estimates of $[\gamma, \theta]$ such that $\tilde{\gamma} \geq 0$. Then,

$$\tilde{\theta} = \overline{Y^R} - \tilde{\gamma} \overline{X^R}.$$

It follows that

$$\tilde{Y}_i^R = \tilde{\gamma} X_i^R + \tilde{\theta} = \tilde{\gamma} X_i^R + (\overline{Y^R} - \tilde{\gamma} \overline{X^R}) = \overline{Y^R} + \tilde{\gamma} (X_i^R - \overline{X^R}), \quad i = 1, \dots, n.$$

Then the sum of squared errors for the prediction of Y^R based on the constrained LS estimates is calculated to be

$$\begin{aligned} \sum_{i=1}^n (Y_i^R - \tilde{Y}_i^R)^2 &= \sum_{i=1}^n \left[(Y_i^R - \overline{Y^R}) - \tilde{\gamma} (X_i^R - \overline{X^R}) \right]^2 \\ &= \sum_{i=1}^n (Y_i^R - \overline{Y^R})^2 + \tilde{\gamma}^2 \sum_{i=1}^n (X_i^R - \overline{X^R})^2 - 2\tilde{\gamma} \sum_{i=1}^n (Y_i^R - \overline{Y^R}) (X_i^R - \overline{X^R}) \\ &= \sum_{i=1}^n (Y_i^R - \overline{Y^R})^2 + \tilde{\gamma}^2 n S_{1,1} - 2\tilde{\gamma} n S_1. \end{aligned}$$

Therefore, in view of (35)-(36),

$$\sum_{i=1}^n (Y_i^R - \tilde{Y}_i^R)^2 \geq \sum_{i=1}^n (Y_i^R - \overline{Y^R})^2, \quad (37)$$

and “=” holds if and only if $\tilde{\gamma} = 0$. This completes the proof. \square

A.4 Proof of Corollary 6

Proof. Under Assumption 1,

$$\Sigma_{\mathbf{X}^R} \rightarrow \text{diag} \{ \text{Var}(X_1^R), \dots, \text{Var}(X_p^R) \} \quad a.s.,$$

and

$$\Sigma_{\mathbf{X}^R, \mathbf{Y}^R} \rightarrow [\text{Cov}(X_1^R, Y^R), \dots, \text{Cov}(X_p^R, Y^R)]^T \quad a.s.. \quad (38)$$

It follows that

$$\hat{\Gamma} \rightarrow \left[\frac{\text{Var}(X_1^R)}{\text{Cov}(X_1^R, Y^R)}, \dots, \frac{\text{Var}(X_p^R)}{\text{Cov}(X_p^R, Y^R)} \right]^T \quad a.s.. \quad (39)$$

Equations (38) and (39) together imply

$$\hat{\gamma}_j S_j \rightarrow \text{Var}(X_j^R) \quad a.s., \quad j = 1, \dots, p,$$

and therefore,

$$P(\hat{\gamma}_j S_j > 0) \rightarrow 1, \quad \text{as } n \rightarrow \infty, \quad j = 1, \dots, p.$$

Hence, if $S_j > 0$,

$$P(\hat{\gamma}_j > 0) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

\square

A.5 Proof of Theorem 3

Proof. Notice that

$$\begin{aligned} E \left[\left(Y_i^R - \hat{Y}_i^R \right) \hat{Y}_i^R \right] &= E \left\{ E \left[\left(Y_i^R - \hat{Y}_i^R \right) \hat{Y}_i^R | X_i^R \right] \right\} \\ &= E \left[\hat{Y}_i^R E \left(Y_i^R - \hat{Y}_i^R | X_i^R \right) \right] \\ &= 0. \end{aligned}$$

Therefore,

$$E \left(Y_i^R \right)^2 = E \left(Y_i^R - \hat{Y}_i^R \right)^2 + E \left(\hat{Y}_i^R \right)^2.$$

This together with the fact that $E \left(\hat{Y}_i^R \right) = E \left(Y_i^R \right)$ yields

$$E \left(Y_i^R - \hat{Y}_i^R \right)^2 = E \left(Y_i^R \right)^2 - E \left(\hat{Y}_i^R \right)^2 = \text{Var} \left(Y_i^R \right) - \text{Var} \left(\hat{Y}_i^R \right). \quad (40)$$

Separately,

$$E \left(Y_i^R - \hat{Y}_i^R \right)^2 = E \left(\epsilon_i^R \right)^2 = E \left(\epsilon_i^U - \epsilon_i^L \right)^2 = 2\sigma^2. \quad (41)$$

By Markov's inequality, we have

$$P \left(\hat{Y}_i^R < 0 \right) \leq P \left(|\hat{Y}_i^R - Y_i^R| > Y_i^R \right) \leq \frac{E \left(Y_i^R - \hat{Y}_i^R \right)^2}{\left(Y_i^R \right)^2}. \quad (42)$$

(42) together with (40) and (41) proves the desired result. \square

References

- [Artstein(1975)] Artstein, Z., Vitale, R.A. (1975). A strong law of large numbers for random compact sets. *Annals of Probability*, 5, 879–882
- [Aumann(1965)] Aumann, R.J. (1965). Integrals of set-valued functions. *J. Math. Anal. Appl.*, 12, 1–12.
- [Billard(2000)] Billard, L., Diday, E. (2000). Regression analysis for interval-valued data. In *Data Analysis, Classification and Related Methods*, Proceedings of the Seventh Conference of the International Federation of Classification Societies (IFCS'00) (pp. 369–374). Springer, Belgium.
- [Billard(2002)] Billard, L., Diday, E. (2002). Symbolic regression analysis. In *Classification, Clustering and Data Analysis*, Proceedings of the Eighth Conference of the International Federation of Classification Societies (IFCS'02) (pp. 281–288). Springer, Poland.
- [Billard(2007)] Billard, L. (2007). Dependencies and variation components of symbolic interval-valued data. In *Selected Contributions in Data Analysis and Classification* (pp. 3–12). Springer, Berlin Heidelberg.
- [Blanco(2011)] Blanco-Fernández, A., Corral, N., González-Rodríguez, G. (2011). Estimation of a flexible simple linear model for interval data based on set arithmetic. *Computational Statistics and Data Analysis*, 55, 2568–2578.
- [Blanco(2012)] Blanco-Fernández, A., Colubi, A., González-Rodríguez, G. (2012). Confidence sets in a linear regression model for interval data. *Journal of Statistical Planning and Inference*, 142, 1320–1329.
- [Carvalho(2004)] Carvalho, F.A.T., Neto, E.A.L., Tenorio, C.P. (2004). A new method to fit a linear regression model for interval-valued data. *Lecture Notes in Computer Sciences*, 3238, 295–306.
- [Cattaneo(2012)] Cattaneo, M.E.G.V., Wiencierz, A. (2012). Likelihood-based imprecise regression. *International Journal of Approximate Reasoning*, 53, 1137–1154.

- [Diamond(1990)] Diamond, P. (1990). Least squares fitting of compact set-valued data. *J. Math. Anal. Appl.*, 147, 531–544.
- [Gil(2001)] Gil, M.A., Lopez, M.T., Lubiano, M.A., Montenegro, M. (2001). Regression and correlation analyses of a linear relation between random intervals. *Test*, 10, 1 183–201.
- [Gil(2002)] Gil, M.A., Lubiano, M.A., Montenegro, M., Lopez, M.T. (2002). Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika*, 56, 97–111.
- [Gil(2007)] Gil, M.A., González-Rodríguez, G., Colubi, A., and Montenegro, M. (2007). Testing linear independence in linear models with interval-valued data. *Computational Statistics & Data Analysis*, 51, 3002–3015.
- [G-R(2007)] González-Rodríguez, G., Blanco, A., Corral, N., and Colubi, A. (2007). Least squares estimation of linear regression models for convex compact random sets. *Advances in Data Analysis and Classification*, 1, 67–81.
- [Hormander(1954)] Hörmander, H. (1954). Sur la fonction d’appui des ensembles convexes dans un espace localement convexe. *Arkiv för Mat*, 3, 181–186.
- [Kendall(1974)] Kendall, D.G. (1974). Foundations of a theory of random sets. In: *Stochastic Geometry*, eds. Harding, E.F. and Kendall, D.G., John Wiley & Sons, New York.
- [Korner(1998)] Körner, R., Näther, W. (1998). Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates. *Information Sciences*, 109, 95–118.
- [Lyashenko(1982)] Lyashenko, N.N. (1982). Limit theorem for sums of independent compact random subsets of Euclidean space. *Journal of Soviet Mathematics*, 20, 2187–2196.
- [Lyashenko(1983)] Lyashenko, N.N. (1983). Statistics of random compacts in Euclidean space. *Journal of Soviet Mathematics*, 21, 76–92.
- [Manski(2002)] Manski, C.F., Tamer, T. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70, 519–546.
- [Matheron(1975)] Matheron, G. (1975) *Random Sets and Integral Geometry*. John Wiley & Sons, New York.
- [Molchanov(2005)] Molchanov, I. (2005) *Theory of Random Sets*. Springer-Verlag, London.
- [Neto(2008)] Neto, E.A.L., Carvalho, F.A.T. (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, 52, 1500–1515.
- [Neto(2010)] Neto E.A., Carvalho, F.A.T. (2008). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, 54, 333–347.
- [Radstrom(1952)] Rådström, H. (1952). An embedding theorem for spaces of convex sets. *Proc. Amer. Math. Soc.*, 3, 165–169.
- [Seber(1977)] Seber, G.A.F. (1977). Linear regression analysis. John Wiley & Sons, New York.
- [Stoyan(1998)] Stoyan, D. (1998). Random sets: models and statistics. *International Statistical Review*, 66, 1, 1-27.
- [Sun(2014)] Sun, Y. and Li, C. (2014). On linear regression for interval-valued data in $\mathcal{K}_C(\mathbb{R})$. preprint. arXiv: 1401.1831.